CMPU 250 Preliminary Analysis Writeup

The movie business is a big business. According to Box Office Mojo, the US domestic box office gross for 2025 has already exceeded $1.6 billion; each of the two preceding years, the total annual gross fell between $8.5 and $9 billion [Box Office Mojo]. These numbers exclude international box office revenues, money spent to physically or digitally purchase or rent movies, and money spent to watch movies through subscription streaming services. In an environment where various new releases compete for the biggest slice of the box office pie, film studios and researchers are acutely interested in predicting which films will be financially successful and orienting release calendars accordingly.[1]

In the early 2000s, various research explored the effect of critical reviews on film success, and at least one study (Basuroy et al.) found that "each of the first eight weeks, both positive and negative reviews are significantly correlated with box office revenue" [King, 2007; Basuroy et al., 2003]. However, with the rise of the internet, consumers no longer turn to critics like Roger Ebert or Pauline Kael when deciding whether to see a film; instead, word-of-mouth (WOM) information on social media sites and crowd-sourced movie ratings like those from IMDb are among the first information presented to prospective filmgoers. Recent research has examined the effect of online WOM on movies and analyzed online review data, including using sentiment analysis to measure online response to a film [Sharma et al., 2023; Yoon et al., 2017; Oghina et al., 2012].

A fundamental difference between crowdsourced and critical movie evaluation is the inherent manipulability of the former. Online communities, working in concert, can shape online movie scores by leaving masses of (usually negative) reviews to lower a movie's score, a phenomenon known as "review bombing." Some recent research has examined this phenomenon, including papers by Schuff et al. (2024) and Tomaselli et al. (2021). Review bombing is most evident in the scores of films which are perceived as having made "woke" an existing (white and/or male) movie or franchise by rebooting it with women and/or people of color in the main roles (i.e. *Ghostbusters* (2016), *The Little Mermaid* (2023)) and films which address controversial topics like the Israel-Palestine conflict or the Armenian Genocide (i.e. *Israelism* (2023), *The Promise* (2016)). The prevalence of review bombing has led sites like IMDb to weight their ratings, leading to sometimes significant gaps between unweighted means and the publicly presented score.[2]

When online reviews can shape film success and by extension future studio decisions, unchecked review bombing has the potential to deter studios and directors from future projects which center or positively portray marginalized groups or address political issues. Ratings from IMDb, which is owned by Amazon, are explicitly displayed on the Amazon Prime streaming service; this represents a direct way in which IMDb ratings contribute to film success, as

[1] Research by Lash & Zhao (2016) and Lee et al. (2016), among others, has explored the use of machine learning techniques to predict movie success and profitability.

[2] 5.2 vs 7.2, 5.2 vs 6.8, 5.8 vs 7.7, and 6.0 vs 6.1 for each of the previously mentioned movies, respectively. The smaller gap for *The Promise* is due to the occurrence of simultaneous *positive* review bombing, a phenomenon also visible on a smaller scale for some of the other films.

higher-rated films are more attractive to users and likely to be shown more often. Other movie recommendation algorithms also rely on some form of score to produce recommendations, although these may be more opaque; review bombing, if not addressed, can directly perpetuate representational and political harms by damaging film's performance in recommendation algorithms.

With this in mind, weighting systems like IMDb's represent direct bulwarks against these harms, and how successful it is at identifying and negating review bombing is critical in determining whether those perpetuating the review bombing are successful in damaging the film and its creators. Our project explores review bombing, score weighting, and film recommendations, analyzing which kinds of films are more likely to review bombed, how that relates to their success, how platforms like IMDb use weighting to minimize the effects of review bombing, and more. In addition to our observational analysis, we plan to create our own film recommendation algorithm, and run it using weighted and unweighted ratings to explore how effective score weighting might be in counteracting review bombing, and whether there are any apparent issues with IMDb's weighting methods.

Research Questions:

1. In terms of genre and keywords, what kinds of movies does IMDb's weighting algorithm boost or de-boost?
2. How do different movie rating weightings affect the rate at which movie recommendation algorithms recommend certain movies?
3. What kind of weighting algorithm would help create more unbiased reviews by offsetting the effects of review-bombing?

We expect to find that controversial movies contain a bimodal distribution, centering around extremes of 1 and 10 when compared to non-controversial movies, suggesting that many users of IMDb rate these controversial movies based on their stance on the controversy rather than the merit of the movie. Furthermore, we predict that the user ratings of politically charged movies will have a greater amount of polarized ratings based on the stance that the user has on the controversy rather than the cinematic quality of the film. We also predict that a more diverse cast, especially ones with more underrepresented groups such as an all-women cast, will draw more polarized ratings; which in return will lead to these movies being deboosted in recommendation algorithms.

We will measure the polarization of rating distributions using two measures, first applied to movie reviews by Hickey and Mehta at FiveThirtyEight in 2017 [Mehta and Hickey (2017)]: Esteban and Ray's polarization score and the proportion of "polarized" votes, ones that are either "1" or "10." We will measure the correlation between movie recommendability and polarized rating distributions using linear regression.

We expect to find that movies belonging to genres that may be politically controversial ("History," "Biography," and "Documentary," for example) will be more likely to have polarized rating distributions and to have average ratings that differ the most across different countries.

Examples of individual movies that support this hypothesis mentioned in the introduction are *The Promise* (2016) and *Israelism* (2023) (though *Israelism* will not be included in our initial analysis because it has fewer than 50,000 user votes). We also expect to find that movies with higher proportions of female and non-white stars will be more likely to have polarized rating distributions. Examples of individual movies that support this hypothesis mentioned in the introduction are *Ghostbusters* (2016) and *The Little Mermaid* (2023).

DATA:

We use a dataset scraped from IMDb's website and GraphQL API, cleaned into imdb-cleaned.csv, to examine how user ratings reflect controversy and polarization. Each row in the dataset represents a single movie and includes variables such as the film's title, genre(s), release year, keywords, rating breakdown (number of votes for scores 1–10), and the five countries with the most user votes. Additional variables were created in the cleaning process to better capture polarization and rating behavior, including unweighted_rating, total_votes, rating_diff (the difference between the IMDb weighted rating and the unweighted average), and polarization_score, based on Esteban and Ray's polarization index. The dataset also tracks the difference in average scores between countries (country_ptp) and identifies which countries contributed the highest and lowest ratings (max_country_ptp, min_country_ptp).

Some important terms it is necessary to define for our project are as follows: polarization, polarization score, international polarization, weighted rating, unweighted rating, 2-9 rating. Some of these are fairly simple; the *weighted rating* is a film's basic overall rating provided by IMDb, while the *unweighted rating* is the unweighted mean rating provided on a film's IMDb page. *2-9 rating* refers to the unweighted mean rating of a film when extreme scores (1s and 10s) are removed and only scores 2-9 are considered. *Polarization* refers to the amount of disagreement in a film's ratings; a film is more polarized if it has a higher *polarization score*. We calculate a film's polarization score using a version of the polarization formula created by Esteban and Ray (1994); the maximum score a film can receive is 1, meaning that exactly half the ratings are 10 and the other half are 1, and the lowest is 0, if all the ratings are identical (Mehta and Hickey (2017) also use a version of this formula, but a slightly different one). *International polarization* of a film between two countries refers to the difference in unweighted mean rating for the two countries. These are some of the basic metrics we will use to identify films that have been review bombed and analyze the extent and nature of the review bombing.

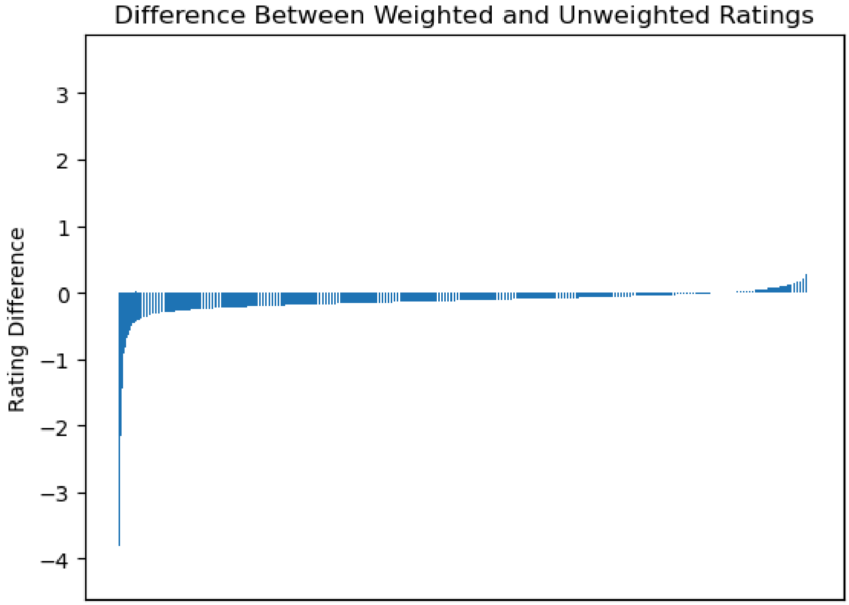
METHODS:

To facilitate our exploration of how review bombing may impact film recommendation systems, we plan to create a naive recommendation algorithm of our own, which takes a film (or potentially multiple films) and recommends similar movies based on content and rating. For each film in the database, IMDb provides a list of "plot keywords," words which pertain to or describe the content of the film. To generate recommendations based on a film (the "base film"), our algorithm will look at the top 50 keywords for the base film and identify potential recommendations by seeing which films have the highest percentage of shared keywords in their own top 50 (or, if there are less than 50 keywords, in the entire keyword list). We will then calculate the percentage of keywords a potential recommendation shares with the base film as a number out of 10, where 0 is no shared keywords and 10 is all keywords in common. For each potential recommendation, that number will then be multiplied by that film's rating to determine a recommendation score designed to capture both film similarity and popularity (for example, a film with 20/50 shared keywords (4) and a rating of 6.5 would receive a score of 26, while a film with 17/50 shared keywords (3.4) and a rating of 8.1 would receive a score of 27.54). This will yield recommendation scores between 0 (no shared keywords, any rating) and 100 (identical keywords, 10 rating). Then, the films with the highest recommendation scores with regard to the base film will be displayed as recommendations.

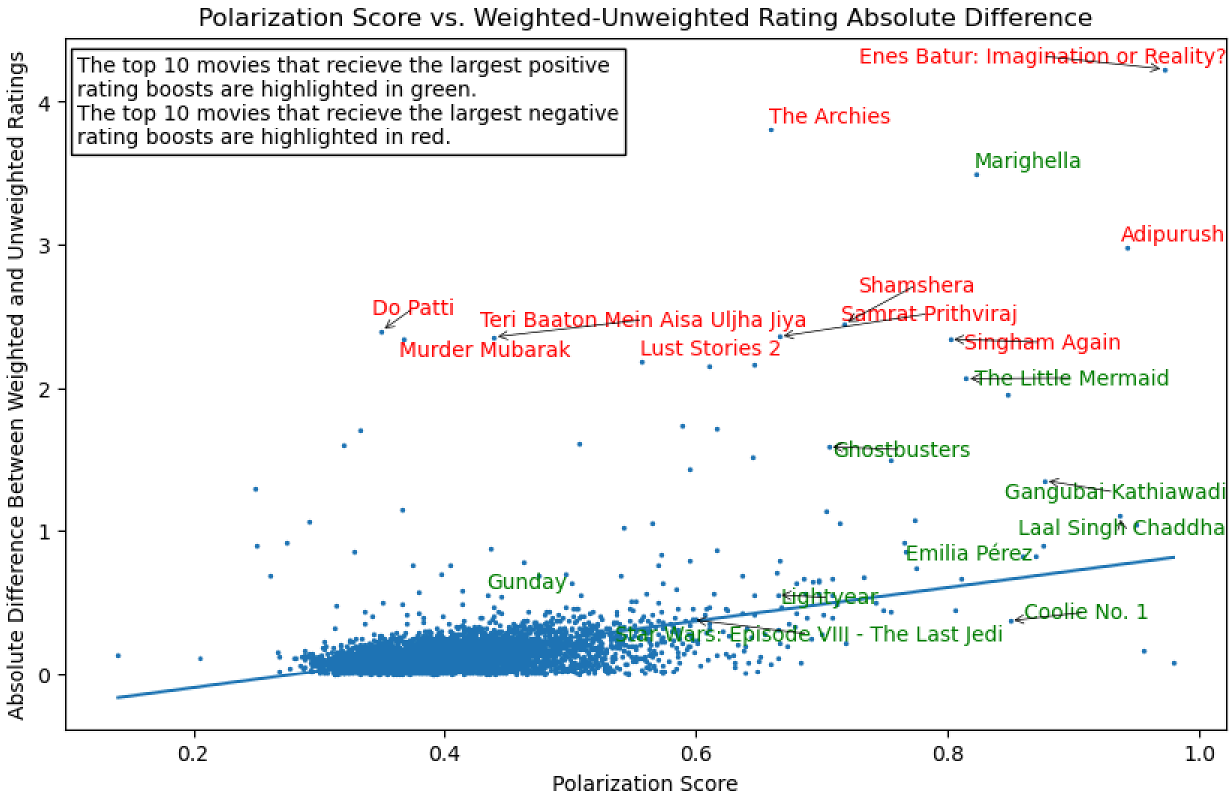
By running the recommendation system with both weighted and unweighted ratings and comparing the difference in results, we aim to illuminate the potential impact of score weighting on recommendation systems. Although our system represents an extremely simplified version of what an actual streaming service's recommendation system might look like, we believe that it should allow for worthwhile analysis of review bombing and score weighting's respective potential effects on recommendation.

RESULTS:

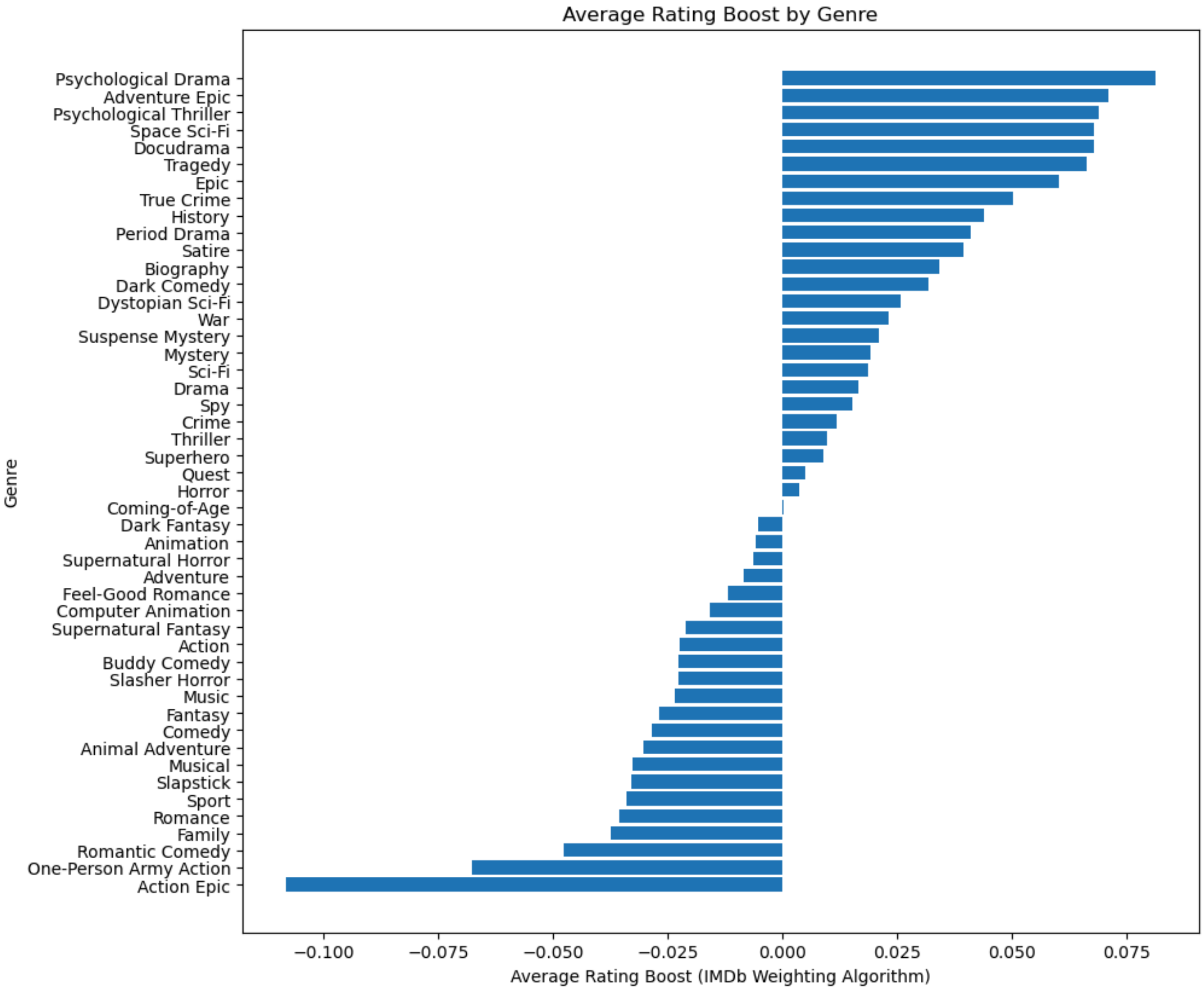
Exploring IMDB's Rating Algorithm



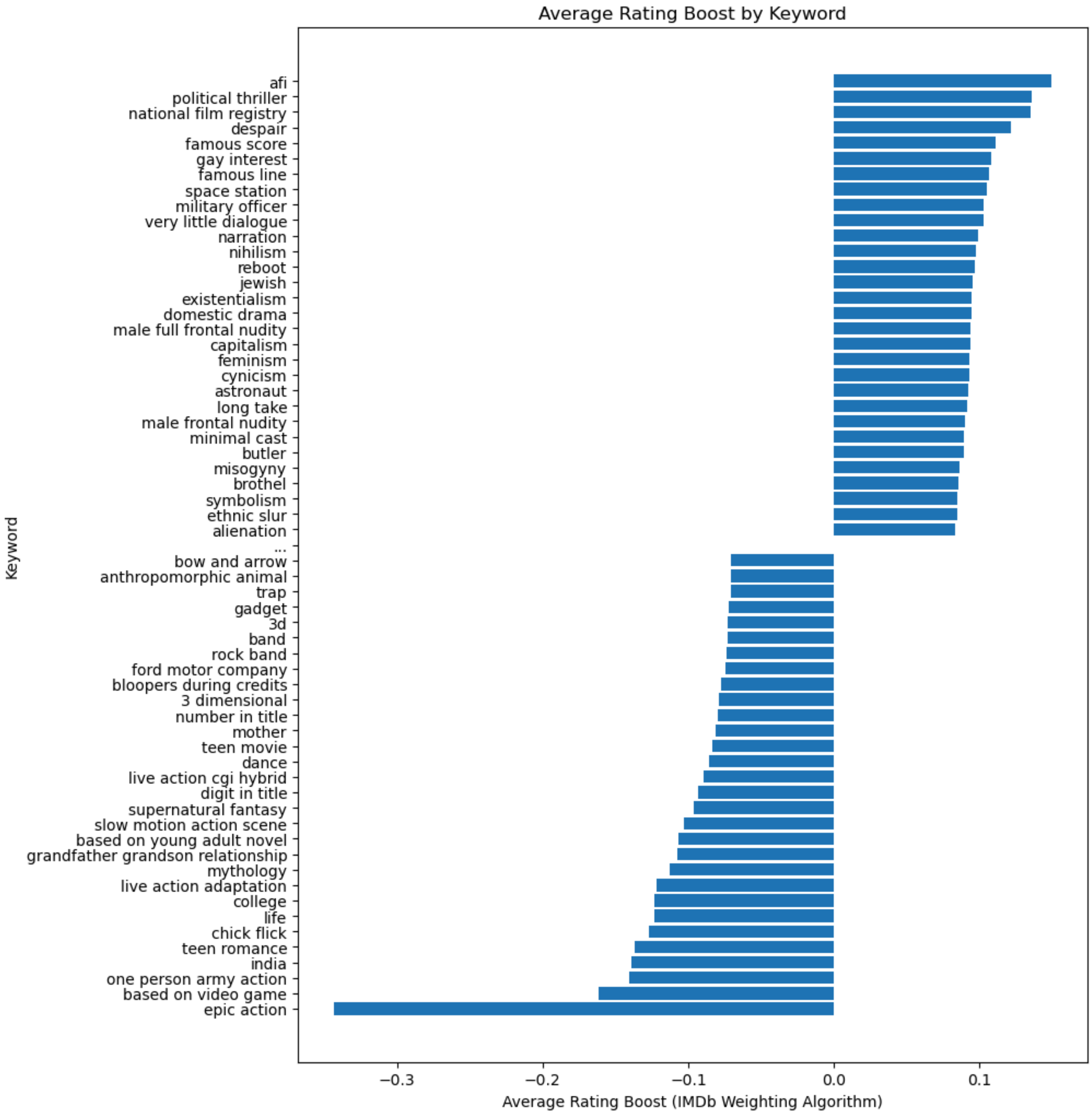
The distribution of rating difference shows that IMDb's weighting algorithm tends to yield lower scores than the unweighted ratings, with some exceptions. Most films' ratings are adjusted by tenths of a point or less, although some, visible on the ends of the distribution, have their ratings drastically altered by the algorithm.



In general, it seems as though films with higher polarization scores are more likely to have their ratings heavily adjusted by IMDb's weighting algorithm (positive, albeit weak, correlation between absolute rating difference and polarization score). This supports the idea that the weight algorithm, either by design or not, mitigates the effects of review-bombing, which results in abnormally polarized rating distributions.

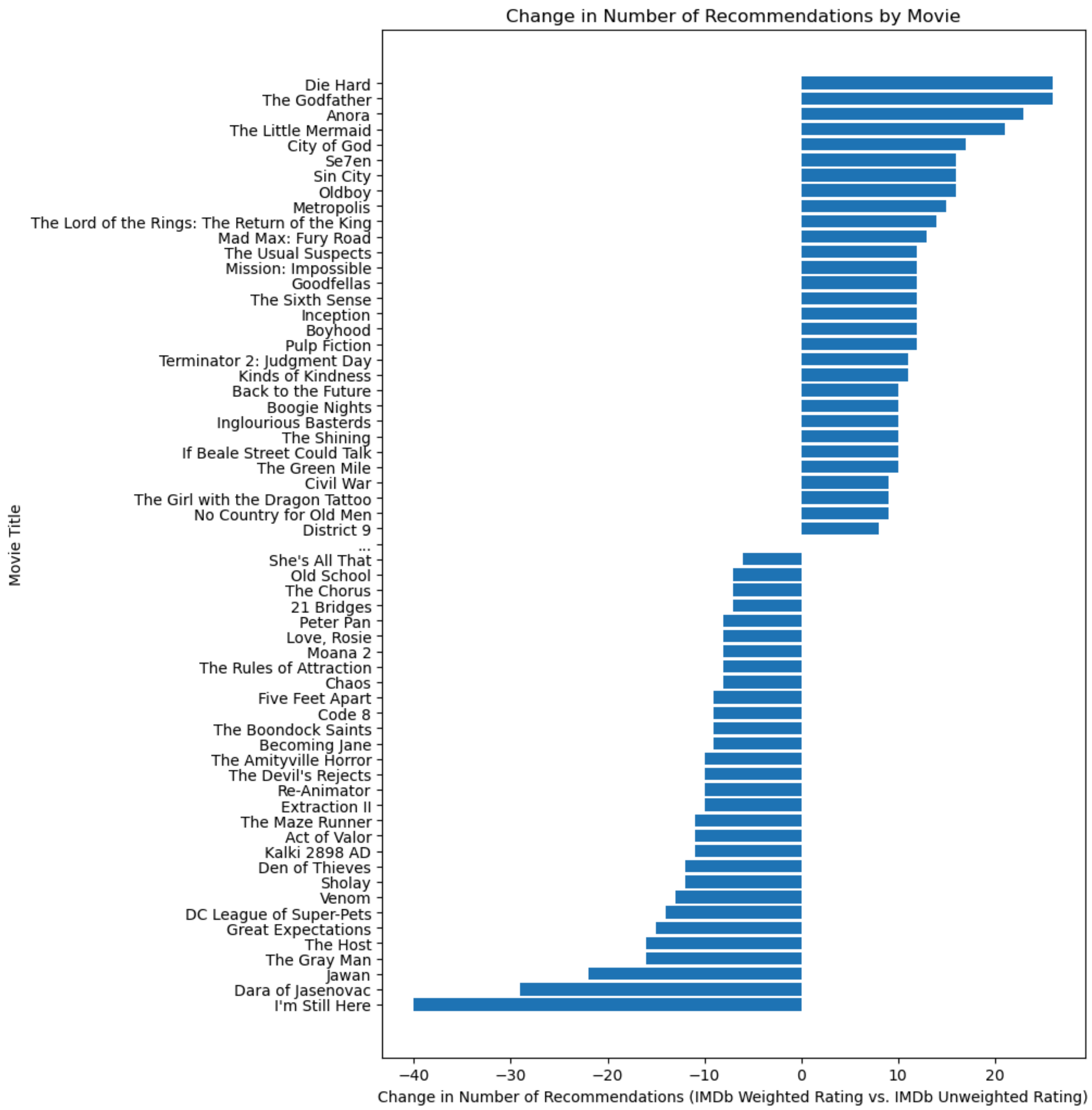


Genres associated with more politically controversial subject material (e.g. "War", "Period Drama", "Docudrama", "History", "Biography", "Satire") as well as genres more likely to have violent or non-culturally normative themes (e.g. "Psychological Drama", "Psychological Thriller", "True Crime", "Dark Comedy", "Dystopian Sci-fi") are among those boosted most by IMDb's weighting algorithm. Genres associated with more mainstream, family-friendly, or culturally normative themes (e.g. "Family", "Romance", "Musical", "Romantic Drama", "Slapstick", "Sport", "Action Epic", "Animal Adventure", "Comedy", "Fantasy") are among those de-boosted most by the algorithm. This suggests that the algorithm is more likely to boost ratings for films with more controversial subject matter, while de-boosting ratings for films with more mainstream or family-friendly themes. This is consistent with the idea that IMDb's weighting algorithm is designed to reduce the influence of unrepresentative user ratings on the overall rating, which may be more likely to be positive for films with more mainstream or family-friendly themes and more negative for films with more controversial subject matter.



Keywords associated with more high-brow movies (e.g. "afi", "national film registry", "very little dialogue", "symbolism", "minimal cast") as well as will controversial subject material (e.g. "political thriller", "gay interest", "jewish", "capitalism", "feminism", "male frontal nudity", "misogyny", "brothel", "ethnic slur", "alienation") are among those boosted most by IMDb's weighting algorithm. Keywords associated with more mainstream movies ("epic action", "based on video game", "teen romance", "chick flick", "college", "live action adaptation", "based on young adult novel", "teen movie") are among those de-boosted most by IMDb's weighting algorithm. This matches the observations on how weighting affects different genres.

Effect of Weighting on a Hypothetical Recommendation Algorithm


Change in Number of Recommendations by Movie
Movie Title
Die Hard
The Godfather
Anora
The Little Mermaid
City of God
Se7en
Sin City
Oldboy
Metropolis
The Lord of the Rings: The Return of the King
Mad Max: Fury Road
The Usual Suspects
Mission: Impossible
Goodfellas
The Sixth Sense
Inception
Boyhood
Pulp Fiction
Terminator 2: Judgment Day
Kinds of Kindness
Back to the Future
Boogie Nights
Inglourious Basterds
The Shining
If Beale Street Could Talk
The Green Mile
Civil War
The Girl with the Dragon Tattoo
No Country for Old Men
District 9
...
She's All That
Old School
The Chorus
21 Bridges
Peter Pan
Love, Rosie
Moana 2
The Rules of Attraction
Chaos
Five Feet Apart
Code 8
The Boondock Saints
Becoming Jane
The Amityville Horror
The Devil's Rejects
Re-Animator
Extraction II
The Maze Runner
Act of Valor
Kalki 2898 AD
Den of Thieves
Sholay
Venom
DC League of Super-Pets
Great Expectations
The Host
The Gray Man
Jawan
Dara of Jasenovac
I'm Still Here
−40
−30
−20
−10
0
10
20
Change in Number of Recommendations (IMDb Weighted Rating vs. IMDb Unweighted Rating)

DISCUSSION: